

ABSTRACT OF THE DISCLOSURE

[0104] Methods, systems, and articles of manufacture consistent with certain principles related to the present invention enable a computing system to perform hierarchical topical clustering of text data based on statistical modeling of co-occurrences of (document, word) pairs. The computing system may be configured to receive a collection of documents, each document including a plurality of words, and perform a modified deterministic annealing Expectation-Maximization (EM) process on the collection to produce a softly assigned hierarchy of nodes. The process may involve assigning documents and document fragments to multiple nodes in the hierarchy based on words included in the documents, such that a document may be assigned to any ancestor node included in the hierarchy, thus eliminating the hard assignment of documents in the hierarchy.

FOOTNOTES

LAW OFFICES

FINNEGAN, HENDERSON,
FARABOW, GARRETT,
& DUNNER, L.L.P.
1300 I STREET, N. W.
WASHINGTON, DC 20005
202-408-4000